DRAFT
2010 TRECVID MULTIMEDIA EVENT DETECTION EVALUATION PLAN

1 Overview

This document presents the evaluation plan for Multimedia Event Detection (MED) track for the TRECVID 2010 evaluation. The multi-year goal of MED is to assemble core detection technologies into a system that can quickly and accurately search a multimedia collection for user-defined events. An event for MED is "*an activity-centered happening that involves people engaged in process-driven actions with other people and/or objects at a specific place and time*".

A user searching for events in multimedia material may be interested in a wide variety of potential events. Since it is an intractable task to build special purpose detectors for each event a priori, a technology is needed that can take as input a definition of the event that a human can use to search a collection of multimedia clips. The MED evaluation series will define events via an **event kit** which consists of:

- An **event name** which is an mnemonic title for the event.
- An **event definition** which is a textual definition of the event.
- An **evidential description** which is a textual listing of the attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it is not a exhaustive list nor is it to be interpreted as required evidence.
- A set of 5-10 **illustrative video examples** each containing an instance of the event. The examples are illustrative in the sense they help form the definition of the event but they do not demonstrate all the inherent variability or potential realizations.

Deleted: prescriptive

Deleted: description of the event in terms of the scenes, objects, people, and activities that provides enough detail for a knowledgeable person to implement the task themselves.

The following topics are discussed below:

- Video source data
- The evaluation task
- Evaluation measures
- Evaluation Infrastructure
- Schedule

2 Video Source Data

The video corpus for the evaluation track is currently being constructed therefore the content below describes the target data resources.

The video source data will consist of user generated content video clips posted on the world wide web. The Linguistic Data Consortium will be the distribution point for the corpus. See the MED '10 web site, https://author.nist.gov/itl/iad/mig/med10.cfm, for licensing and acquisition instructions. The provided resources will include video clips, MED event annotations, and ancillary metadata for each clip.

The target size of the corpora for MED '10 is 100 hours of clips which will be evenly divided into

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010

a 50-hour development (*MEDdev10*) corpus and a 50-hour evaluation (*MEDeval10*) corpus. Developers may use the MEDdev10 corpus in any manner to build their systems, including activities such as dividing it into internal test sets, jack-knifed training, etc. During the summer months, NIST will conduct a dry run evaluation using the MEDdev10 corpus. While testing on the development data is a non-blind test, the purpose of the dry run is to test the evaluation infrastructure which is most easily accomplished using the development data.

Both the MEDdev10 and MEDeval10 corpora will be released early in the evaluation cycle to give people the opportunity to preprocess the full corpus throughout the summer. The evaluation set must not be inspected or mined for information until after the annotations are released for the evaluation set. However, participants can run feature extraction programs on the MEDeval10 set to prepare for the formal evaluation.

Allowable side information (i.e., "contextual" information) will be provided in CSV (comma separated values) data tables. The exact content of the side information is TBD.

3 Evaluation Task

The MED task is: given an Event Kit, find all clips that contain the event in a pre-indexed metadata store of the search corpus.

Deleted: which defines an event

The MED task is a "multimedia" task in that systems will be expected to detect evidence of the event using either or both the audio and video streams of the clips. The events used for the MED '10 evaluation can be found on the MED '10 web site. Participant may implement systems for one or all of the specified events.

4 Evaluation Infrastructure

Systems will be evaluated on how well they can detect MED event instances in the evaluation corpus. The determination of correct detection will be at the clip level, i.e. systems will provide a response for each clip in the evaluation corpus. For testing purposes, each event will be considered independent

System detection performance is measured as a tradeoff between two error types: missed detections (MD) and false alarms (FA). The two error types will be combined into a single error measure using the Normalized Detection Cost (NDC) model, which is a linear combination of the two errors. The NDC model distills the needs of an application profile into a set of predefined constant parameters that include the event priors and weights for each error type. The single operation point characterized by the NDC model is a small window into the performance of an event detection system. In addition to NDC measures, Detection Error Tradeoff (DET) curves [2] will be produced to graphically depict the tradeoff of the two error types over a wide range of operational points. The NDC model and the DET curve are related: the NDC model defines an optimal point along the DET curve.

The rest of this section defines the system output, followed by the two steps of the evaluation process: Decision Error Tradeoff (DET) curve production, and NDC computations.

4.1 System Outputs

Systems will record system outputs in a to be specified CSV file format. The system will

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010

generated the following data for each clip for each event:

- **Decision score**: A numeric score indicating how likely the event observation exists with more positive values indicating more likely observations.
- **Actual Decision**: A Boolean value indicating whether or not the event observation should be counted for the primary metric computation.

The decision scores and actual decisions permit performance assessment over a wide range of operating points. The decision scores provide the information needed to construct the DET curve. The actual decisions provide the mechanism for the system to indicate which putative observations to include in the NDC calculation: i.e., the putative decisions with a *true* actual decision.

Systems must ensure their decision scores have the following two characteristics: first, the values must form a non-uniform density function so that the relative evidential strength between two putative terms is discernable. Second, the density function must be consistent across events for a single system so that event-averaged measures using decision scores are meaningful.

Since the decision scores are consistent across events, the system must use a single threshold for differentiating *true* and *false* actual decisions.

4.2 Detection Error Tradeoff Curves

Graphical performance assessment uses a Detection Error Tradeoff (DET) curve that plots the system's missed detection probabilities (P_{Miss}) and false alarm probabilities (P_{FA}) that are a function of a detection threshold, Θ. This Θ is applied to the system's detection scores meaning the clips with decision scores above the Θ are 'declared' to be the set of detected instances. After Θ is applied, the following measurements are then computed separately for each event. The per-event formulas for P_{Miss} and P_{FA} are:

$$P_{Miss}(S, E_i, \Theta) = \frac{N_{Miss}(S, E_i, \Theta)}{N_{Targ}(E_i)}$$

$$P_{FA}(S, E_i, \Theta) = \frac{N_{FA}(S, E_i, \Theta)}{N_{NonTarg}(E_i)}$$

Where:

$N_{Miss}(S, E_i, \Theta) = number\ of\ missed\ detections\ for\ system\ S, event\ E_i\ at\ decision\ score\ \Theta$

$N_{Target}(E_i) = number\ of\ clips\ containing\ event\ instances\ for\ event\ E_i$

$N_{NonTarg}(E_i) = number\ of\ clips\ that\ do\ not\ contain\ event\ instances\ for\ event\ E_i$

$N_{FA}(S, E_i, \Theta) = number\ of\ false\ alarms\ for\ event\ E_i at\ decision\ score\ \Theta$

4.3 DCR Computations

The evaluation will use the Normalized Detection Cost (*NDC*) measure for evaluating system performance. *NDC* is a weighted linear combination of the system's probabilities of Missed Detection and False Alarm. The measure's derivation can be found in Appendix A and the final formula is summarized below. NIST will report an NDC for each event and not average them over events.

$$NDC(S, E_i, \Theta) = Cost_{Miss} \cdot P_{Miss}(S, E_i, \Theta) \cdot P_{Targ} + Cost_{FA} \cdot P_{FA}(S, E_i, \Theta) \cdot (1 - P_{Targ})$$

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010

Where:

$$E_i = \text{the } i^{th} \text{ event}$$
$$Cost_{Miss} = \boldsymbol{TBD}: a\ constant\ defining\ the\ cost\ of\ a\ missed\ detections.$$
$$Cost_{FA} = \boldsymbol{TBD}: a\ constant\ defining\ the\ cost\ of\ a\ false\ alarm.$$
$$P_{Target} = \boldsymbol{TBD}: a\ constant\ defining\ the\ a\ priori\ rate\ of\ event\ instances.$$

The measure's unit is in terms of cost per clip used. NDC has been normalized so that an *NDC*=0 indicates perfect performance and an *NDC*=1 is the cost of a system that provides no output, i.e. $P_{Miss}=1$ and $P_{FA}=0$.

Two versions of the NDC will be calculated for each system: the Actual NDC and the Minimum NDC.

4.3.1 Actual NDC

The Actual NDC is the primary evaluation metric. It is computed by counting clips with *true* actual decisions as clips the system declares to contain the event.

4.3.2 Minimum DCR

The Minimum NDC is a diagnostic metric. It is found by searching the DET curve for the Θ with the minimum NDC. The difference between the value of Minimum NDC and Actual NDC indicates the benefit a system could have gained by selecting a better threshold.

5 Submission of results

Submissions to NIST will be required only to allow NIST to perform a system-mediated improvements to the test set ground truth.

Submissions will be made via ftp according to the instructions in Appendix B. In addition to the system output, NIST requests a system description be supplied for each submission. This description should include: a description of the hardware used to process the data, computational resources (cpu runtime, memory footprint, etc.) and a description of the architecture and algorithms used in the system such as the features or reasoning process.

6 Schedule

Consult the main schedule on the TREVID 2010 web site http://www-nlpir.nist.gov/projects/tv2010/#schedule.

7 References

[1] Harold W. Kuhn, "The Hungarian Method for the assignment problem", *Naval Research Logistic Quarterly*, **2**:83-97, 1955.

[2] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", Eurospeech 1997, pp 1895-1898.

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010

Appendix A: Derivation of Normalized Detection Cost

Normalized Detection Cost (*NDC*) is a weighted linear combination of the system's Missed Detection and False Alarm probabilities. The constant parameters of NDC, which are specified below, represent both the richness of events in the source data and the relative detriment of particular clip detection errors to a hypothetical application.

The cost of a system begins with the cost of missing an event ($Cost_{Miss}$) and the cost of falsely detecting an event ($Cost_{FA}$). $N_{Miss}(S,E)$ is the number of missed detections for system *S*, event *E*. $N_{FA}(S,E)$ is the number of false alarms for the same system and event.

$$DetectionCost(S,E) = Cost_{Miss} \cdot N_{Miss}(S,E) + Cost_{FA} \cdot N_{FA}(S,E)$$

To facilitate comparisons across systems and test sets, we divide Detection Cost by the number of video clips N_{Trials}.

$$\begin{aligned} DetectionCost(S,E) &= \frac{\mathrm{Cost_{Miss}} \cdot N_{Miss}(S,E) + Cost_{FA} \cdot N_{FA}(S,E)}{\mathrm{N_{Trials}}} \\ &= \mathrm{Cost_{Miss}} \cdot \frac{\mathrm{N_{Miss}(S,E)}}{\mathrm{N_{Trials}}} + \mathrm{Cost_{FA}} \cdot \frac{\mathrm{N_{FA}(S,E)}}{\mathrm{T_{Trials}}} \\ &= \mathrm{Cost_{Miss}} \cdot \frac{\mathrm{N_{Miss}(S,E)}}{\mathrm{N_{Targ}(E)}} \cdot \frac{\mathrm{N_{Targ}(E)}}{\mathrm{N_{Trials}}} + \mathrm{Cost_{FA}} \cdot \frac{\mathrm{N_{FA}(S,E)}}{\mathrm{N_{NonTargTrials}}} \cdot \frac{\mathrm{N_{NonTargTrials}(S,E)}}{\mathrm{N_{Trials}}} \\ &= Cost_{Miss} \cdot P_{Miss}(S,E) \cdot P_{Target}(E) + Cost_{FA} \cdot R_{FA}(S,E) \cdot (1 - P_{Target}(E)) \end{aligned}$$

$P_{Target}(E)$ is the probability of a clip containing the event. This value is dependent on the event but providing this prior to a system for each event changes the definition of an event – it includes the event definition and the prior. Instead, we replace the event-dependent prior with a single, global prior, P_{Target}, that in combination with the $Cost_{Miss}$ and $Cost_{FA}$ reflects the characteristics of an application profile. Since the evaluation corpus will have an engineered richness, the single prior is warranted. The modified formula becomes:

$$DetectionCost(S,E) = Cost_{Miss} \cdot P_{Miss}(\mathrm{S,E}) \cdot P_{Target} + Cost_{FA} \cdot P_{FA}(\mathrm{S,E}) \cdot (1 - P_{FA}(\mathrm{S,E}))$$

The range of the DCR_{Sys} measure is [0,∞). To ground the costs, a second normalization scales the cost to be 0 for perfect performance and 1 to be the cost of a system that provides no output (either providing no output, $P_{Miss} = 1$ and $P_{FA} = 0$, or declaring every clip to be an instance). The resulting formula is the Normalized Detection Cost of a system (*NDC*).

$$NormDectectionCost(S,E) = \frac{DetectionCostRate(S,E)}{MINIMUM(Cost_{Miss} \cdot P_{Target}, Cost_{FA} \cdot (1 - P_{Target}))}$$

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010

Appendix B: Submission Instructions

This section is TBD.

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010

Appendix C: Corpus File Naming Conventions

This section is TBD.

Deleted: MED10-EvalPlan-V01

Deleted: May 7, 2010